Class attendance, peer similarity, and academic performance in a large field study

Kassarnig, Valentin; Bjerre-Nielsen, Andreas; Mones, Enys; Lehmann, Sune; Lassen, David Dreyer

Published in:
PLOS ONE

DOI:
10.1371/journal.pone.0187078

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

PLOS ONE

RESEARCH ARTICLE

Class attendance, peer similarity, and academic performance in a large field study

Valentin Kassarnig[1]*, Andreas Bjerre-Nielsen[2,3], Enys Mones[4], Sune Lehmann[2,4,5], David Dreyer Lassen[2,3]

1 Institute of Software Technology, Graz University of Technology, Graz, Austria, **2** Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark, **3** Department of Economics, University of Copenhagen, Copenhagen, Denmark, **4** Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark, **5** The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

* kassarnig@ist.tugraz.at

Citation: Kassarnig V, Bjerre-Nielsen A, Mones E, Lehmann S, Lassen DD (2017) Class attendance, peer similarity, and academic performance in a large field study. PLoS ONE 12(11): e0187078. https://doi.org/10.1371/journal.pone.0187078

Editor: Paula B. Andrade, Universidade do Porto, Faculdade de Farmácia, PORTUGAL

Received: April 21, 2017

Accepted: September 24, 2017

Published: November 8, 2017

Data Availability Statement: The entire Copenhagen Networks Study, including the current study, has been approved by the Danish Data Protection Agency (DDPA), which is the relevant legal entity in Denmark. The data used in this study includes third-party reported information on grades obtained from administrative sources. According to the Act on Processing of Personal Data, such data cannot be made available in the public domain. We confirm that the data is available upon request to all interested researchers under conditions stipulated by the DDPA. Data inquiries should be addressed

Abstract

Identifying the factors that determine academic performance is an essential part of educational research. Existing research indicates that class attendance is a useful predictor of subsequent course achievements. The majority of the literature is, however, based on surveys and self-reports, methods which have well-known systematic biases that lead to limitations on conclusions and generalizability as well as being costly to implement. Here we propose a novel method for measuring class attendance that overcomes these limitations by using location and bluetooth data collected from smartphone sensors. Based on measured attendance data of nearly 1,000 undergraduate students, we demonstrate that early and consistent class attendance strongly correlates with academic performance. In addition, our novel dataset allows us to determine that attendance among social peers was substantially correlated (>0.5), suggesting either an important peer effect or homophily with respect to attendance.

Introduction

An increasing number of individuals seek a high level of education to secure their future and improve their economic possibilities [1]. Academic performance is an essential factor in the success of the post-education period with respect to employment [2]. For this reason, the ability to predict students' academic success has been the subject of increasing interest. The knowledge regarding expected academic performance is also a valuable input for educators and school administrators, as this information can be used to identify and target vulnerable students at risk of dropping out or in need of additional attention. However, gathering information about attendance levels using conventional methods (surveys or self-reports) is subject to inherent biases [3] and moreover, can be costly to gather at the scale of schools or universities.

Here we propose a new method for measuring attendance. This new methodology overcomes important limitations of previous approaches. Specifically, our method leverages data collected via smartphone sensors to identify class locations from clusters of students following

to the Social Fabric steering committee, to be reached at ddl@econ.ku.dk.

Funding: This work was supported by the Villum Foundation, the Danish Council for Independent Research, and University of Copenhagen (via the UCPH-2016 grant Social Fabric and The Center for Social Data Science).

Competing interests: The authors have declared that no competing interests exist.

the same courses and estimate then the students' attendance (see Methods for details). We used the measured attendance levels of almost 1,000 university students to investigate the relationship between students' attendance and their grades, as well as the social aspects of academic performance. This is the first time a dataset of comparable richness has been used to conduct analyses on attendance.

The theoretical literature on student achievement emphasizes that class attendance is associated with better performance. One strand of theoretical literature are the pedagogical models where class attendance can be seen as student involvement, among other features which also highlights the resources of the school and the content being taught [4]. Other theoretical approaches include economic models where rational agents decide on optimal usage of time spent studying vs. leisure/other courses [5].

There is a large body of existing data-driven research on class attendance, absenteeism and their impact on academic achievements [6–23] as well as on the relationship between behavior of peers [24–28]. However, the methodology applied in the large majority of previous work has limitations: results are based on analyzing surveys, sign-in-sheets or other types of self-reports, which are known to be prone to biases and errors [3]. Two exceptions that collected data from sensors are the *StudentLife* study [23] and Zhou et al. [29]. The *StudentLife* study used location data recorded on students' phones [23]. Students were considered to be at class when they spent at least 90% of the scheduled period at the class location. Although there were variations observed in class attendance, they found no relation between final grades and absence (either initial level or the pace of absence over the term) which contradicts the findings of most related studies. A likely explanation of the observations in the *StudentLife* study is the small sample size (< 50 students). Another approach is used by Zhou et al. [29] who employ connectivity data from the WiFi network at Tsinghua University. The location of students and consequently their class attendance was determined by studying how students' phones connected to the nearly 2,800 WiFi access points with known locations distributed over university campus. Based on nine weeks of observations, Zhou et al. found that students with higher GPAs attended classes at a higher rate. Moreover, compared to low performing students, they were also found to be more likely to arrive late to class. Our approach shares some similarities with [23, 29] but there are also some fundamental differences as we have also investigated dynamic patterns (i.e. early and consistent attendance throughout the semester) and considered the social environment.

Most existing studies have found that class attendance is a significant and positive predictor of course grades [6–20, 22]. More specifically, the meta study by Crede et al. [20] concludes that attendance is the most accurate known predictor of academic performance, superior to scores on standardized admissions tests such as the SAT, high school GPA, study habits, and study skills. Some studies report on experiments which have quantified the importance of attendance on exam performance through mandatory attendance [17] and intentionally omitting parts of the curriculum [19], and both found a significant effect on the number of exam questions answered correctly. In addition to general attendance throughout the semester, initial attendance has been shown to be an important predictor of academic success [7]. Previous results also indicate that average attendance drops over the course of the semester, irrespective of performance [10, 21, 30].

The connection between attendance and peer behavior has been explored to a lesser extent. Only a few preliminary observations exist in the literature based on co-occurrence at class or workplace [24–28]. One example is the work of De Paola, where individual absence behavior was found be related to peer group absenteeism [28]. Yet these studies are limited by having no access to data regarding actual communication or interactions.

Educational policies aimed at increasing attendance have a broad set of tools. The standard approach is to enact mandatory attendance as experimented in [17]. Other tools include tutoring [31] and nudging [32]. Moreover, for children and adolescents other options are to involve partnerships from schools to parents and their communities [33, 34].

The aim of our study was to evaluate the accuracy of measuring class attendance from smartphone data and assess its usefulness for discovering new patterns in attendance. These aims were divided into three specific objectives:

1. To what extent could students' class attendance be inferred from data collected by smartphones? As a specific question we investigated whether a method based on finding large clusters of students enrolled in the same course was sufficient.
2. How accurate was the measure of students' attendance at predicting their subsequent exam performance?
3. What is the degree of similarity in attendance between students who are social peers?

Materials and methods

Data

We employed data collected in the Copenhagen Networks Study (CNS) [35]. As part of the CNS project, various data types were recorded by dedicated smartphones from nearly 1,000 undergraduate students at the Technical University of Denmark (DTU), over a period of 2 years. The channels for data collection include location (using GPS), proximity of other students (via Bluetooth scans) and mobile phone communication. The CNS covered the academic years 2013/14 and 2014/15 which form the basis of our analysis. About 78% of the sampled students were male and 22% female. They were divided in 24 different study lines (majors) and more than 60% of them were newly enrolled in 2013; more than 25% in 2012; the remaining enrolled in 2011. Their course grades and schedules were provided by the Technical University of Denmark.

At the university, course attendance is non-mandatory and is not logged for the offered courses. The educational model used in classes at the Technical University of Denmark is quite varied, ranging from lectures coupled with classroom exercises to more modern forms, e.g. flipped class room teaching or characterized by problem-based group work. Also note that due to the possibility of exiting the experiment at any given point, the number of participants varied over time.

In terms of privacy all participating students have provided full consent for use of their data for research purposes. At any time a student could exit the study and request to have their data deleted. The collected data has been authorized by Danish Data Protection Agency. For further details of the CNS project and a short overview of the obtained data please refer to [35].

Class location and attendance

We first needed to calculate the locations of the classes before we could use them to determine the attendance of each eligible student (see Fig 1). The list of course participants and the associated class times were retrieved from the list of course grades and the course page archive, respectively. To improve the accuracy of class estimation, we only considered courses with the standard DTU length of 4 hours and with at least 8 participants. A significant fraction of classes did not have a fixed location throughout the 4-hour period as the students may change building or room (e.g. from a lecture hall to a laboratory). Therefore, the designated time of the class was divided into sixteen 15-minute bins and for each time bin a separate class location

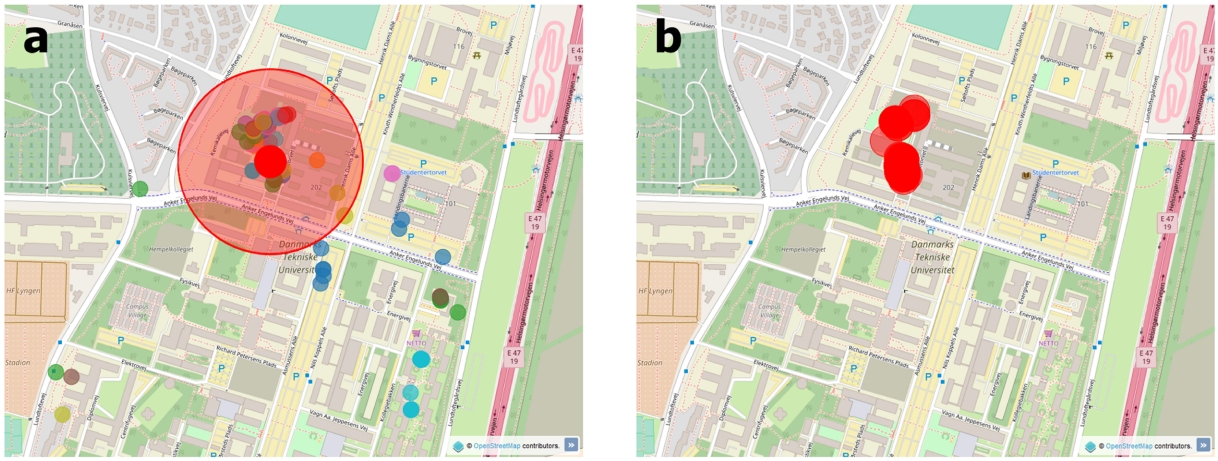

Fig 1. Estimation of class locations at the Technical University of Denmark campus. a) Location of the students who were assigned to a specific class (colored small circles), with the estimated location (moderate size red circle) and range of the class (large light red circle) in a particular day. b) Estimated class locations (red circles) of the same course throughout the semester. (Map data copyrighted OpenStreetMap contributors and available from www.openstreetmap.org under CC BY-SA 2.0).

https://doi.org/10.1371/journal.pone.0187078.g001

was estimated, with our results being robust to changes of bin size. For a particular time bin, we determined the proximity network of students from the Bluetooth scans representing nearby co-students within a distance of 15 m. In this network we identified the primary cluster, represented by the highest degree node (that is, the student surrounded by the most co-students signed up for the same course) and its direct neighbors. In contrast to other procedures of determining the primary cluster (e.g. using the largest connected component), this method is robust against the noisy proximity data because some missing links do not necessarily affect the cluster. Once the main cluster was found, the location of the class was defined by the median location of the members in the cluster, using his/her location with the highest accuracy during the 15-minute period. If the estimated location was inside the university campus, we accept it as the location for the class in the corresponding time bin (shown in Fig 1a). The fact that our method uses Bluetooth data to identify class locations is a significant advantage, since we do not need to rely on official records. The lack of reliance on official records makes our approach applicable even when such records are not available or when the records do not match the actual class locations. This sensor based approach has (to the best of our knowledge) not been used previously.

The attendance of a participant was based on their location relative to the estimated class location in each bin. All students who were no further than 200 m from the estimated class location were assigned to the class in the specific time bin (all results were robust against variations in the distance threshold in the range 5–500 m). The value of 200 m was explained by the noise in the measurement of location, especially when using GPS data inside buildings [36]. Members of the main cluster were automatically assigned to the class. Fig 1b shows the estimated locations for a specific course throughout the semester. For further analyses we only considered students as attending when they were within the 200 m range in at least three time bins to avoid false positives due to accidental proximity to the class location.

We also tested our method against the actual course schedules and their locations. Fig 2 shows the cumulative distribution of error in distance based on more than 26,000 class location estimations. More than 75% of estimated locations were found to be within the range of 100 m of the scheduled location, and a 200 m threshold includes 90% of the classes. Note that the error was estimated using the center of the corresponding building instead of the actual

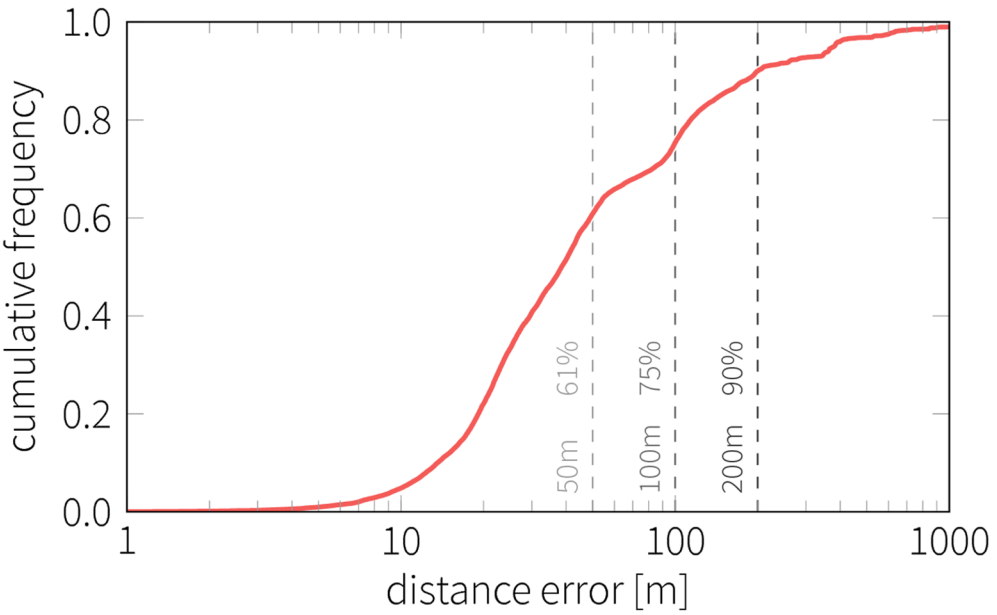

Fig 2. Accuracy of class location inference. The curve shows the cumulative distribution of distance error, that is, the distance between the designated location of the class and our estimation based on the location of the students. Dashed lines mark the distance errors of 50 m, 100 m and 200 m along with the corresponding percentage of classes with error below those thresholds.

https://doi.org/10.1371/journal.pone.0187078.g002

room which caused some imprecision for larger buildings. Furthermore, there were cases of class relocations which did not appear in the official records and therefore the actual error was typically lower than our estimate.

We have implemented an interactive visualization tool accompanying the paper that helps understand our approach. The tool and its source code is available on GitHub: valentin012.github.io/class-loc-d3/.

Social ties

The exchange of text messages between two people suggests a strong social tie [37] and thus, we used this to infer the students' social connections. That is, for each student, the list of their peers was constructed from those participants they had sent a message to or received a message from. For the sake of simplicity and based on the sparsity of the network of text messages, we did not set any lower limit on the number of messages exchanged, that is, a single message is sufficient to establish a link in our network.

Statistical methods

In order to measure the correlation between two observed variables we used Spearman's correlation coefficient. This nonparametric procedure does not assume a linear relation between the two variables since it only tests the association between their ranks instead of their raw values. The coefficient ranges from +1 to −1 indicating a perfect positive or negative correlation, respectively, whereas 0 indicates no correlation.

When comparing the distribution of observed variables from different groups we relied on the Kruskal—Wallis H-test. This nonparametric test examines whether all the observations originate from the same underlying population. As follow-up post-hoc test we used Dunn's

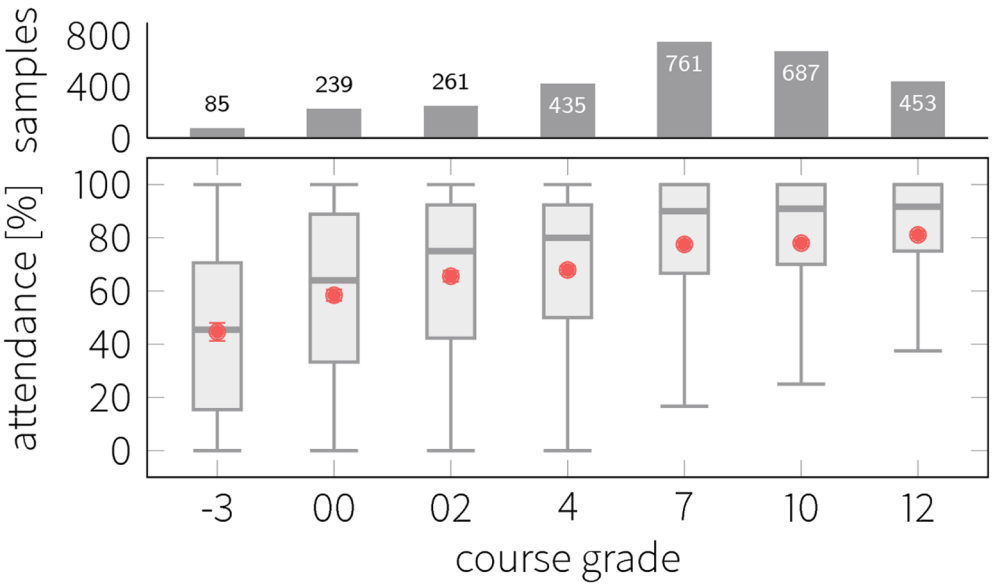

Fig 3. Class attendance conditional on grade obtained. Box plots show median values (solid horizontal line), mean values (red dots), lower and upper quartiles (box outline) and lower and upper fences (quartiles ± IQR, whiskers). Error bars mark standard deviation of the mean. Bar chart above the boxplot shows the number of observations in each grade group.

https://doi.org/10.1371/journal.pone.0187078.g003

multiple comparison test with Bonferroni correction to reveal which groups are significantly different from each other.

In order to observe temporal trends in the data we used a Theil—Sen estimator. This non-parametric line-fitting technique is more robust against outliers and skewed data than, for instance, simple linear regression.

Results

In the following we analyze attendance patterns for a group of students at the Technical University of Denmark (see Methods for an explanation of how attendance is computed). First, we show that attendance is correlated with the achieved grades both at the level of a specific course and overall performance (i.e. average term grade). Second, we look at the temporal aspects and show that there is a general decrease in attendance over the course of a semester regardless of the performance. However, the attendance behavior of low and high performing students displays substantial differences with respect to time. Finally, we show to what extent individual students share similar attendance patterns with their social peers.

Academic achievement

Fig 3 shows the aggregate statistics for all grades (all courses and students considered) as a function of the final grade. Grades follow the Danish grading system that spans between -3 and 12 with 7 distinct grade points (we also denote the corresponding grades according to the US system): -3 and 00 (grade F), 02 (grade D), 4 and 7 (grade C), 10 (grade B) and 12 (grade A). A positive correlation can be observed in Fig 3 with respect to the mean and median, which is quantified by a Spearman correlation of .255 ($p < .001$); the value of the coefficient indicates a weak (< 0.3) positive correlation. Observations where the individual fails to show up at the exam were excluded. The two plots at the boundary of the range (-3 and 12) show distinct, well-separated boxes marking a significant difference in the distributions.

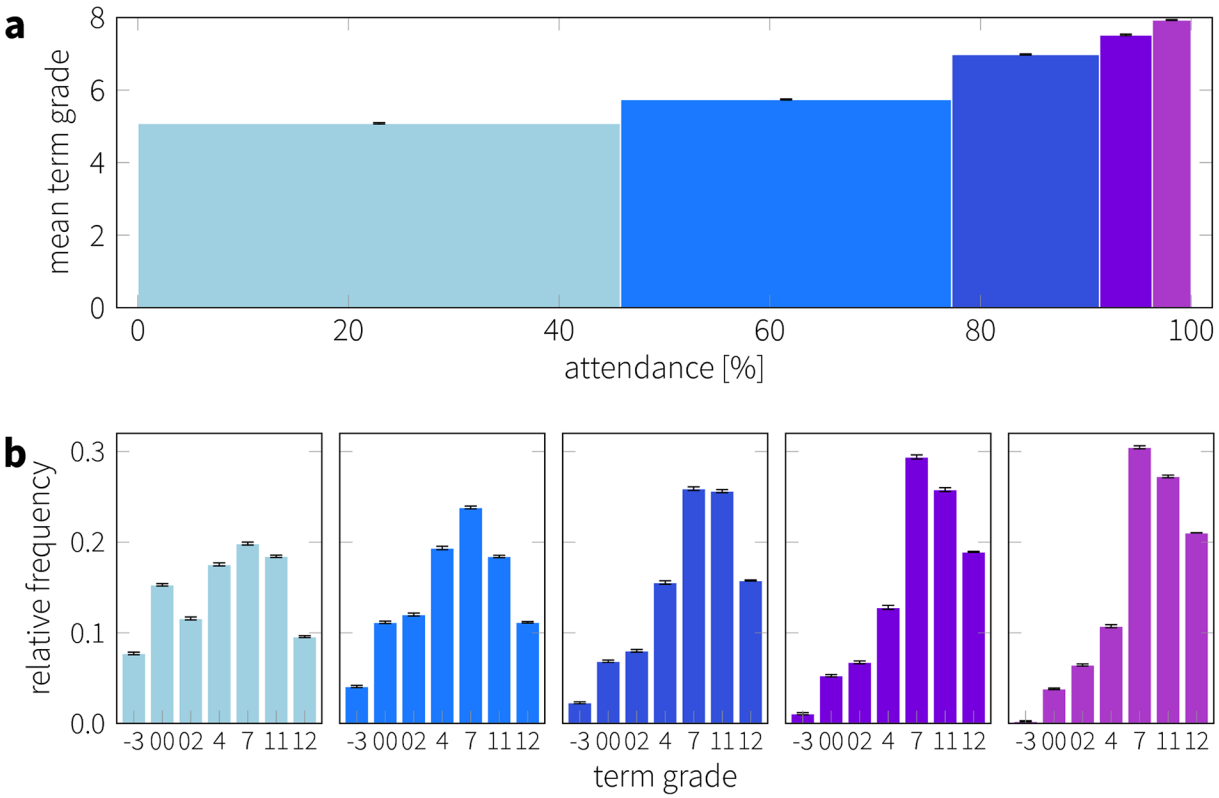

Fig 4. Class attendance and performance. a) Illustration of the five groups based on class attendance. Bars indicate groups of equal size, width corresponds to the span of attendance percentage in the specific group and height shows the mean term GPA. b) Grade distribution inside each attendance quintile. Each group includes at least 373 students.

https://doi.org/10.1371/journal.pone.0187078.g004

To further quantify the observed trend corresponding to the level of attendance, we divided the population into five quintiles based on the students' attendance and measure the performance inside these groups (see Fig 4). The distribution of attendance is illustrated in Fig 4a, where each group is represented by a single bar and the width of the bars is adjusted to span over the covered attendance level. The majority of the students were characterized by a high attendance (60% of the students attend more than 75% of the classes). The height of the bars (mean term grade) show the observed correlation between attendance and performance. The actual distribution of grades also shows variations over the attendance groups, as displayed in Fig 4b. Low attendance (leftmost) results in a broad distribution of grades, indicating that the performance is not solely a function of the attendance but strongly depends on other factors too. However, groups of high attendance (middle to right) develop a peak at grade 7 and were characterized by an increasingly dominating likelihood of high grades. The rate of failing (grade -3 and 00) in exams drops from 23% (leftmost histogram) to less than 4% (rightmost histogram). Also note that there is no observable difference between the grade distribution of the last two attendance groups. The lack of difference suggests that attendance is better at discriminating between whether or not a student is likely to fail rather than predicting the actual grade achieved provided that the student passes the exam; this observation is supported by separate work on the CNS dataset [38], where the predictive power of not only class attendance, but of many other behavioral factors, is considered. To statistically evaluate the variation in the distribution over the groups, we performed a Kruskal—Wallis H-test. This test rejected the global null hypothesis with $p < .001$ that the medians of the groups are all equal. A follow-up

Table 1. Results of Dunn's multiple comparison test with Bonferroni correction for the grade distributions of different attendance groups. The table contains corrected *p*-values for each pairwise comparison, corresponding to the null hypothesis that the pair of groups has equal medians.

	Low att.	L-M att.	Mod att.	M-H att.	High att.
Low att.	-				
L-M att.	.230	-			
Mod att.	<.001	<.001	-		
M-H att.	<.001	<.001	.663	-	
High att.	<.001	<.001	<.001	1.0	-

https://doi.org/10.1371/journal.pone.0187078.t001

Dunn multiple comparison test with Bonferroni correction revealed pair-wise differences among the groups. The recorded *p*-values can be seen in Table 1. All groups separated by at least one quintile were significantly different ($p < .001$), whereas the difference between some neighboring groups (e.g. Mod. vs. M-H, M-H vs. High) could be only confirmed at a lower significance level, further supporting our remarks in Fig 4b. Although attendance accounts for a significant fraction of variation observed in the distribution of grades, it should be noted that this does not necessarily indicate a causal effect.

Temporal effects

Besides differences in attendance across the population, there is also varying attendance for individual students over the duration of the semester (Fig 5). For the sake of simplicity, we divided our observation of student and course into three groups: the first group is characterized by low grades in the course (grades -3, 00 and 02); the second is moderate performance (grades 4 and 7,) and; finally high achievers (grades 10 and 12). With about 41%, moderate performers constituted the largest fraction of nearly 8,400 observations, followed by high (37%) and the low achievers (22%). We computed the average attendance for each group over the semester (see Fig 5a) and observed a general decrease. Further, low performers showed a drop already in the first week with an attendance level 10% lower than that of the high performers. This initial difference increases further throughout the semester as the rate of absence among low performers is consistently higher compared to the moderate and high performer groups. The total drop in the attendance among low performers is above 20% points, compared to the 10% points and 8% points observed among the moderate and high performers. The corresponding Kruskal—Wallis H-tests rejected the global null hypothesis ($p < .001$) that the medians of the groups are all equal. The corresponding Dunn multiple comparison test with Bonferroni correction suggested significant difference between the observations from every pair of performance groups ($p < .05$ between mod. and high performers; $p < .001$ for others). These differences in the trends were further supported by a Theil—Sen estimation for slopes: -1.4% points/week for low performers, opposed to the -.6 and -.4% points/week measured in the moderate and high performer groups. This difference in the slopes is portrayed in the accompanying inset of Fig 5a, where we show the distribution of slopes measured in all pairs of data points in the trends of the main plot. The low performers' distribution of slopes is clearly separated and significantly different from those describing moderate and high performers ($p < .001$ for low vs. mod. and $p < .01$ for low vs. high). Note that differences in the decrease in attendance are not significant between moderate and high performers ($p = 1.0$), also supporting the hypothesis that (the absence of) attendance affects the failing rate to a higher extent than the actual grades. Finally, at the end of the semester, the initial difference inflates to 24% points between low and high performers

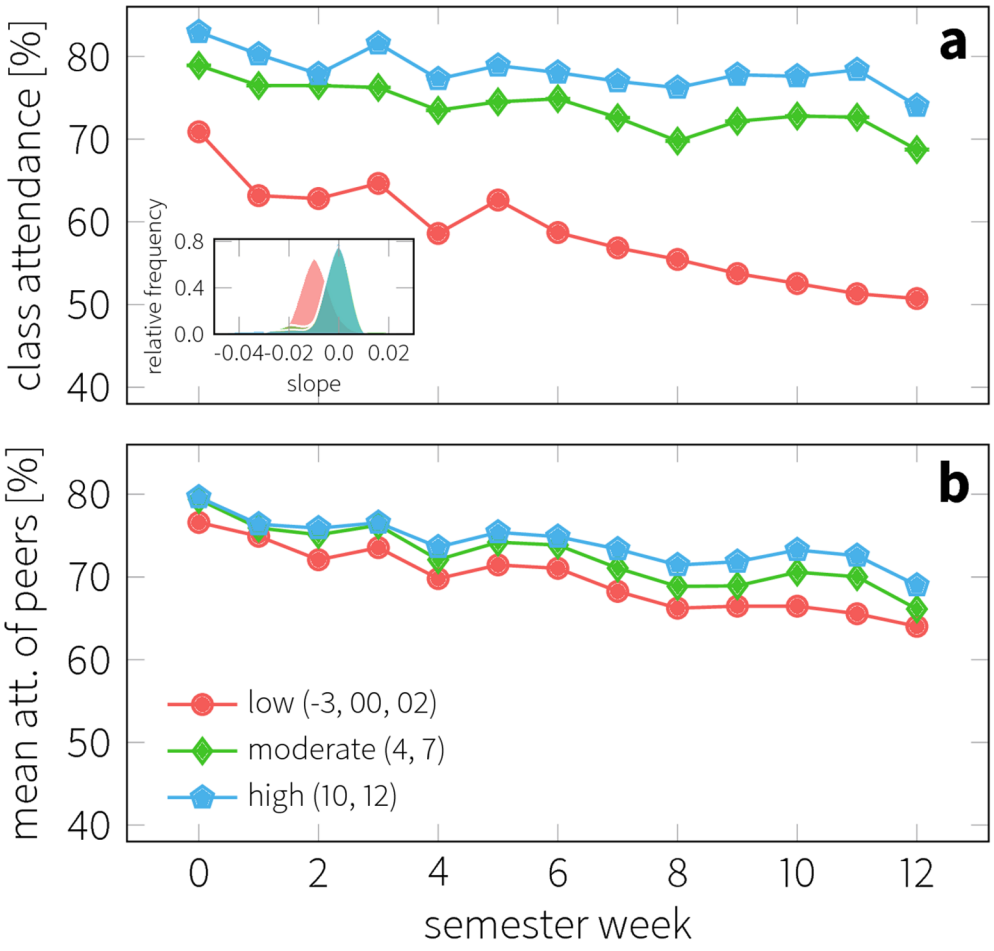

Fig 5. Change in class attendance over a single semester. a) Trends of attendance observed in the three performer groups: low (red circles), moderate (green diamonds) and high performers (blue pentagons), according to the Danish grading system. Inset shows the distribution of slopes measured for each pairs of data point in the trends. b) Mean attendance measured among the contacts of the students based on exchanged text messages.

https://doi.org/10.1371/journal.pone.0187078.g005

due to the faster dropping rate. In summary, the temporal attendance behavior of different types of students differs in the way that low performers start out with a lower attendance rate which also decreases more rapidly throughout the semester compared to that of moderate and high performers.

Peer similarity

Next, we investigated the social aspects of academic performance and attendance: the plot in Fig 5b illustrates the mean attendance level of the peers in the different performance groups. For each individual we created a list of their strongest ties and then calculated the average attendance of them. Strong social contacts were inferred from text message exchanged between two students, as this form of communication indicates a strong bond [37]. On average, each student has exchanged text messages with 4.4 other students from among the participants. Surprisingly, we observed the same differences as above, although less pronounced than in the individual trends: the peers of the low performers also display lower attendance. This supports

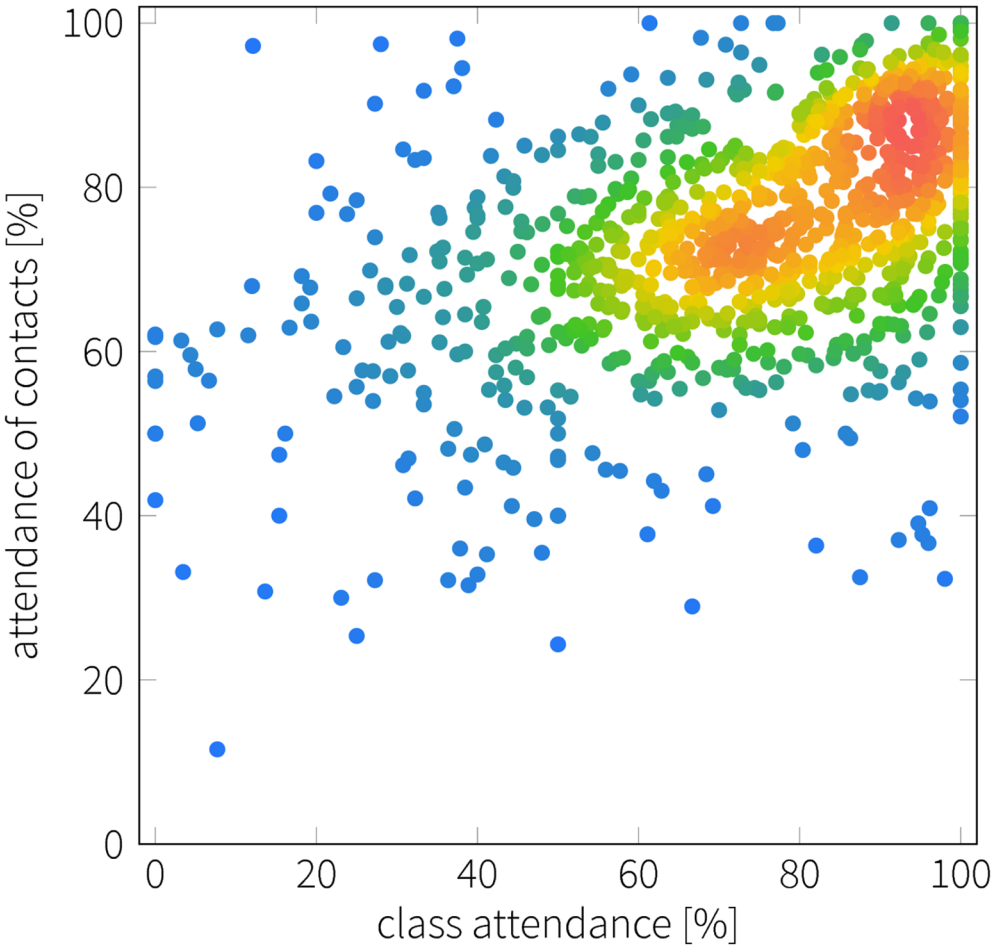

Fig 6. Correlation between own and peers' attendance. Scatter plot that shows student's own attendance vs mean attendance among contacts (inferred from text messages) at the course level. Color represents the relative density of data points.

https://doi.org/10.1371/journal.pone.0187078.g006

previous findings on homophily and peer effects: students communicate more with others who are similar in performance (note that contacts are based on text messages and not on physical proximity).

The observed correlation between own attendance and attendance measured in the ego-networks (the student and their contacts) is clear at the individual level as well. Fig 6 shows the attendance of contacts as a function of the students own attendance, along with the density of observations (color of the dots in plot). Similarity between own and peers' attendance is visible in this scatter plot and has a moderate correlation of .48 ($p < .001$). Furthermore, as shown by the relative density of data points in Fig 6, the number of students is characterized by a peak in peer attendance for high own attendance (above 60%). In other words, peer attendance has a narrow distribution at high attendance levels, compared to the more broad distribution observed below own attendance of 50%. The pattern of similarity is robust against removal of class-level effects as seen in Fig 7. The class effects were removed by subtracting the mean attendance for each class that took place. Robustness of the results suggests that the similarity is driven by homophily or peer effects.

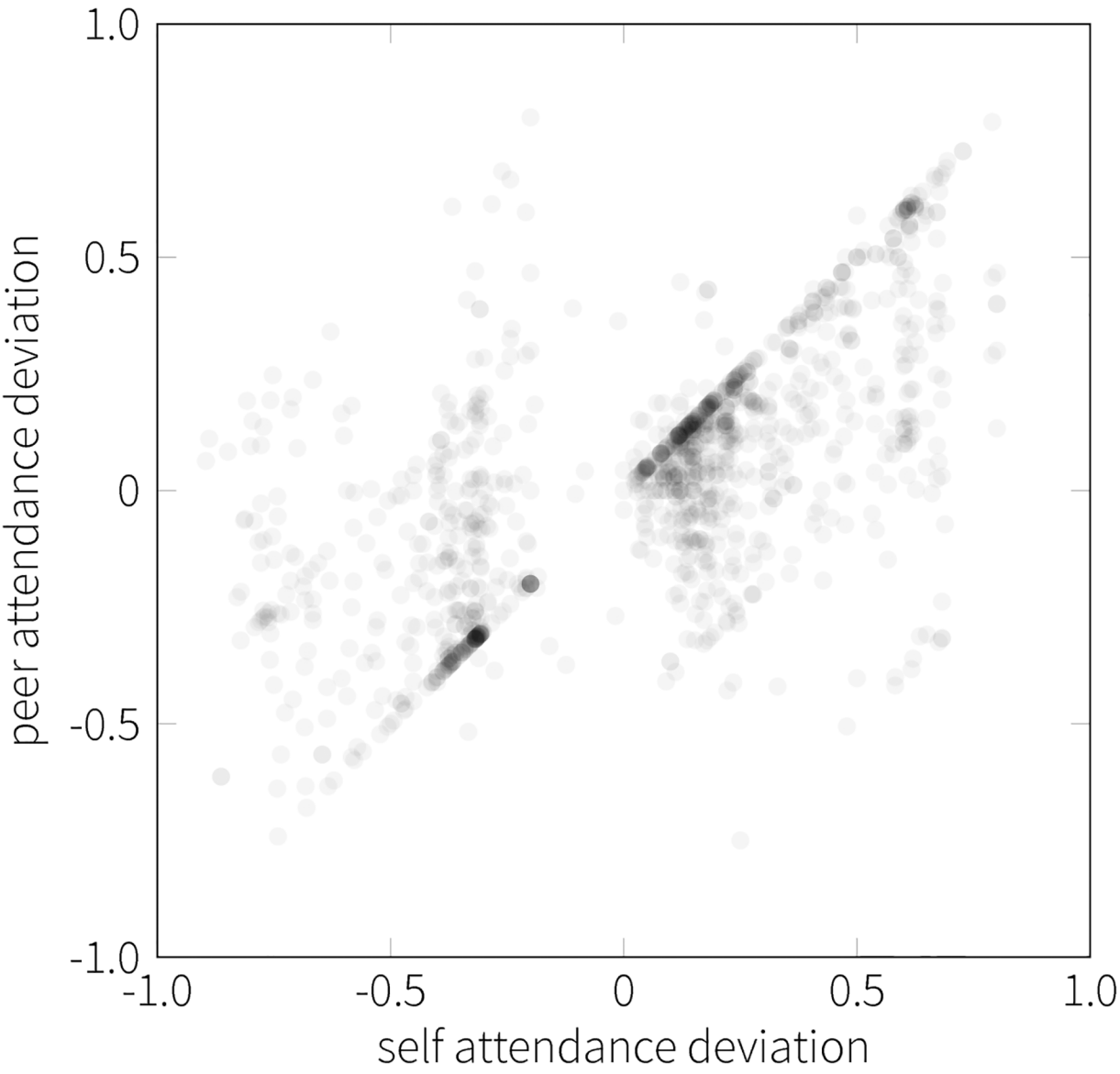

Fig 7. Correlation between self and peer corrected attendance. Shown are the attendance relative to the average calculated at the level of classes. Each dot represents a student's own and their peers' median deviation from the average attendance with respect to a course, that is, their attendance deviation averaged over all classes of a course.

https://doi.org/10.1371/journal.pone.0187078.g007

Discussion

In this paper, we introduced a novel high precision method to measure class attendance in an academic setting. Contrary to previous studies, our method overcomes various limitations caused by restricted data collection techniques. The accuracy of the method makes it possible to scale our measurement and allows for very high precision inference of attendance compared to standard survey-based measurements. Applying our method to a population of nearly 1,000 undergraduate students, we have shown that in this population attendance is not only weakly correlated (< 0.3) with academic success (supporting previous studies) but it is also reflected in the social interactions, which show that students of similar performance tend to be clustered

in the network. Our results suggest a strong mixture of either sorting or peer effects which appears early in the semester and spans across the entire term.

Based on the detailed measurements, we then investigated the temporal aspects of attendance. A general decreasing trend describes the entire population in our dataset, however, there is a clear difference with respect to the pace and level of the effect conditional on performance, indicating a strong early differentiation in behavior between low performers and the rest of the population. Interestingly, this effect vanishes when moderate and high performers are compared. This distinction between low performers and all others is also present in the aggregate statistics of attendance: rate of failing drops by 80% (compared to the value measured in the low attendance quintile) once the attendance reaches the 75% threshold. These results indicate that the effect of attendance on performance shows a complex pattern: while attendance is a strong predictive measure for the failing rate, the effect is less pronounced at high attendance levels and high performance.

Using the contacts obtained from mobile phone communications, we were able to investigate the social aspects of attendance. We found that homophily is present in all groups of academic performance levels, however, it is stronger among high performers. This is supported by an overall robust trend observed in the relative attendance (compared to peers) among high performers, as well as in the correlation between own and peer attendance at the individual level.

We note, however, that our method and the results have some limitations that we address in the following. First, estimation of class locations is based on Bluetooth and GPS signals, both of which are subject to noise. We briefly outline how we mitigate these errors. Our approach is to first identify clusters of physical meetings. We do this by including every Bluetooth signal found between two devices, irrespective of the signal strength, which results in contacts within a typical distance of 10-15 m [39]. Although scans may fail, we reduce the resulting error by employing scans from both devices between two students—this only requires one successful scan and therefore the error should be minimal. Subsequently we employ the identified clusters to estimate the class location. By using the median location of the most connected student within the cluster, we remove noise inherent in location data. After applying our corrections, less than 10% of the estimated class locations are found to be outside of the 200 m radius around the official locations.

Note that it is beyond the scope of this study to control for personal characteristics. However, in [38], attendance is shown to be an important predictor of subsequent performance when controlling for performance of friends and personality. Also, our estimates are not causal estimates of attendance, cf. [17, 19], as the choice of attendance is likely to be correlated with other factors. That is, when a student attends more classes it is not likely to lead to a change in achievements equal to the one we observe. This is due to the fact that attendance and performance could be driven by same latent factor which is both fixed and unobserved to us as researchers.

Another limitation comes from the fact that we have measurements on a subset of all students enrolled in the courses. As an illustration, around 40% of first year students accepted a phone from among the entire cohort of freshmen. The students who participated in the study are different from the average student as they achieve higher grades [40]. We nevertheless observe high variation with respect to attendance and performance within our dataset and, thus, we believe that our results are appropriate precursors of the trends present in the larger student population.

We have demonstrated the connection between attendance and performance; however, attendance alone does not imply active participation. Students who attend class may or may not participate actively in class activities, and although no significant correlation has been

observed e.g. between seating position and performance [7], it is still unknown what other factors contribute to the academic performance. Our methods could be used in conjunction with further experiments such as [17, 19] to yield additional insights on effects of class attendance.

Finally, academic performance is a complex question with multiple facets, and limiting the measurement of success to the raw value of grades is an oversimplification of a high dimensional problem. A first concern is that students were susceptible to different subjects and show interest in distinct fields. A more detailed analysis of performance could, e.g., restrict the analysis above to single subjects. Although this would provide a better understanding of individual performance, the aim of the paper was to investigate the connection at the basic level of attendance. Results with different performance levels suggest that attendance is an efficient predictor of failing, indicating that differentiation at higher orders (that is, among good performers) indeed requires more detailed knowledge regarding the individuals themselves. Further research is therefore needed to understand how factors beyond attendance influence academic success.

Supporting information

S1 File. Dataset details. The supporting information contains further details on the data collection process and the recorded data types. Additionally, we discuss our method on correcting attendance data on the class level and show that we observe distinct attendance-performance correlations for courses of different subject areas.
(PDF)

Author Contributions

Conceptualization: Enys Mones.

Writing – original draft: Valentin Kassarnig, Andreas Bjerre-Nielsen, Enys Mones, Sune Lehmann, David Dreyer Lassen.

Writing – review & editing: Valentin Kassarnig, Andreas Bjerre-Nielsen, Enys Mones, Sune Lehmann, David Dreyer Lassen.

References

1. Carnevale AP, Rose SJ, Cheah B. The College Payoff: Education, Occupations, Lifetime Earnings. Georgetown University Center on Education and the Workforce. 2011.
2. Wise DA. Academic achievement and job performance. The American Economic Review. 1975; 65 (3):350–366.
3. Eagle N, Pentland AS, Lazer D. Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences. 2009; 106(36):15274–15278. https://doi.org/10.1073/pnas.0900282106
4. Astin AW. Student Involvement: A Developmental Theory for Higher Education. Journal of college student personnel. 1984; 25(4): 297–308.
5. Kelley AC. The student as a utility maximizer. The Journal of Economic Education. 1975; 6(2): 82–92. https://doi.org/10.2307/1182457
6. Schmidt RM. Who maximizes what? A study in student time allocation. The American Economic Review. 1983; 73(2):23–28.
7. Buckalew L, Daly JD, Coffield K. Relationship of initial class attendance and seating location to academic performance in psychology classes. Bulletin of the Psychonomic Society. 1986; 24(1):63–64. https://doi.org/10.3758/BF03330504
8. Brocato J. How much does coming to class matter? Some evidence of class attendance and grade performance. Educational Research Quarterly. 1989.

9. Park KH, Kerr PM. Determinants of academic performance: A multinomial logit approach. The Journal of Economic Education. 1990; 21(2):101–101. https://doi.org/10.1080/00220485.1990.10844659
10. Van Blerkom ML. Class attendance in undergraduate courses. The Journal of psychology. 1992; 126 (5):487–494. https://doi.org/10.1080/00223980.1992.10543382
11. Romer D. Do students go to class? Should they? The Journal of Economic Perspectives. 1993; 7 (3):167–174. https://doi.org/10.1257/jep.7.3.167
12. Durden GC, Ellis LV. The effects of attendance on student learning in principles of economics. The American Economic Review. 1995; 85(2): 343–346
13. Devadoss S, Foltz J. Evaluation of factors influencing student class attendance and performance. American Journal of Agricultural Economics. 1996; 78(3):499–507. https://doi.org/10.2307/1243268
14. Gump SE. The cost of cutting class: Attendance as a predictor of success. College Teaching. 2005; 53 (1):21–26. https://doi.org/10.3200/CTCH.53.1.21-26
15. Krohn GA, O'Connor CM. Student effort and performance over the semester. The Journal of Economic Education. 2005; 36(1):3–28. https://doi.org/10.3200/JECE.36.1.3-28
16. Lin TF, Chen J. Cumulative class attendance and exam performance. Applied Economics Letters. 2006; 13(14):937–942. https://doi.org/10.1080/13504850500425733
17. Marburger DR. Does mandatory attendance improve student performance? The Journal of Economic Education. 2006; 37(2):148–155. https://doi.org/10.3200/JECE.37.2.148-155
18. Stanca L. The effects of attendance on academic performance: Panel data evidence for introductory microeconomics. The Journal of Economic Education. 2006; 37(3):251–266. https://doi.org/10.3200/JECE.37.3.251-266
19. Chen J, Lin TF. Class attendance and exam performance: A randomized experiment. The Journal of Economic Education. 2008; 39(3):213–227. https://doi.org/10.3200/JECE.39.3.213-227
20. Credé M, Roch SG, Kieszczynka UM. Class attendance in college a meta-analytic review of the relationship of class attendance with grades and student characteristics. Review of Educational Research. 2010; 80(2):272–295. https://doi.org/10.3102/0034654310362998
21. Nyamapfene A. Does class attendance still matter? engineering education. 2010; 5(1):64–74. https://doi.org/10.11120/ened.2010.05010064
22. Westerman JW, Perez-Batres LA, Coffey BS, Pouder RW. The relationship between undergraduate attendance and performance revisited: Alignment of student and instructor goals. Decision Sciences Journal of Innovative Education. 2011; 9(1):49–67. https://doi.org/10.1111/j.1540-4609.2010.00294.x
23. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2014. p. 3–14.
24. Card D, Giuliano L. Peer effects and multiple equilibria in the risky behavior of friends. Review of Economics and Statistics. 2013; 95(4):1130–1149. https://doi.org/10.1162/REST_a_00340
25. Ejrnæs M, Holm A, Le Maire D. Should I Stay or Should I Go: Peer Effects in Absenteeism. Centre for Applied Microeconometrics—University of Copenhagen; 2014. 3.
26. Winkelmann R. Wages, firm size and absenteeism. Applied Economics Letters. 1999; 6(6):337–341. https://doi.org/10.1080/135048599353032
27. Hesselius P, Nilsson JP, Johansson P. Sick of your colleagues' absence? Journal of the European Economic Association. 2009; 7(2-3):583–594. https://doi.org/10.1162/JEEA.2009.7.2-3.583
28. De Paola M. Absenteeism and peer interaction effects: evidence from an Italian public institute. The Journal of Socio-Economics. 2010; 39(3):420–428. https://doi.org/10.1016/j.socec.2010.02.004
29. Zhou M, Ma M, Zhang Y, SuiA K, Pei D, Moscibroda T. EDUM: classroom education measurements via large-scale WiFi networks. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2016. p. 316–327.
30. Wang R, Harari G, Hao P, Zhou X, Campbell AT. SmartGPA: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2015. p. 295–306.
31. Roderick M, Kelley-Kemple T, Johnson DW and Beechum NO. Preventable Failure: Improvements in Long-Term Outcomes When High Schools Focused on the Ninth Grade Year. University of Chicago Consortium on Chicago School Research; 2014.
32. Rogers T, Duncan T and Wolford T, Ternovski J, Subramanyam S and Reitano, A. A Randomized Experiment Using Absenteeism Information to "Nudge" Attendance. Regional Educational Laboratory Mid-Atlantic, no. 252; 2017.

33. Epstein JL and Sheldon SB. Present and accounted for: Improving student attendance through family and community involvement. The Journal of Educational Research. 2002; 95(5):308–318. https://doi.org/10.1080/00220670209596604

34. Sheldon SB. Improving student attendance with school, family, and community partnerships. The Journal of Educational Research. 2007; 100(5):267–275. https://doi.org/10.3200/JOER.100.5.267-275

35. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring large-scale social networks with high resolution. PloS one. 2014; 9(4):e95978. https://doi.org/10.1371/journal.pone.0095978 PMID: 24770359

36. Moen R, Pastor J, Cohen Y. Accuracy of GPS telemetry collar locations with differential correction. The Journal of Wildlife Management. 1997; p. 530–539. https://doi.org/10.2307/3802612

37. Van Cleemput K. "I'll see you on IM, text, or call you": A social network approach of adolescents' use of communication media. Bulletin of Science, Technology & Society. 2010; 30(2):75–85. https://doi.org/10.1177/0270467610363143

38. Kassarnig V, Mones E, Bjerre-Nielsen A, Sapiezynski P, Lassen DD, Lehmann S. Academic Performance and Behavioral Patterns; 2017. arXiv:1706.09245.

39. Sekara V, Lehmann S. The Strength of Friendship Ties in Proximity Sensor Data. PLoS ONE. 2014; 9(7):e100915. https://doi.org/10.1371/journal.pone.0100915 PMID: 24999984

40. Bjerre-Nielsen A, Dreyer Lassen D. Opportunity and Similarity in Dynamic Friendships; 2017.